DOCUMENT RESUME

ED 178 589 TM 009 779

AUTHOR Mitchell, Blythe C.
TITLE A Glossary of Measurement Terms. Test Service Notebook, 13.
INSTITUTION Harcourt Brace Jovanovich, Inc., New York, N.Y. Test Dept.
NOTE 9p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Testing; *Measurement; *Psychological Testing; *Statistical Analysis; Testing; Vocabulary

ABSTRACT

This glossary of terms used in educational and psychological measurement is geared toward the non-specialist in measurement. Terms which are considered basic are defined, such as those used in test manuals or educational journals. Definitions are based on current usage in textbooks in the areas of educational and psychological measurement and statistics. Approximately 100 terms are defined. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED178589

Test Service Notebook 13

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Marilyn Lucas
Psychological Corp
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A Glossary of Measurement Terms

BLYTHE C. MITCHELL, *Consultant, Test Department*

This glossary of terms used in educational and psychological measurement is primarily for persons with limited training in measurement, rather than for the specialist. The terms defined are the more common or basic ones such as occur in test manuals and educational journals. In the definitions, certain technicalities and niceties of usage have been sacrificed for the sake of brevity and, it is hoped, clarity.

The definitions are based on the usage of the various terms as given in the current textbooks in educational and psychological measurement and statistics, and in certain specialized dictionaries. Where there is not complete uniformity among writers in the measurement field with respect to the meaning of a term, either these variations are noted or the definition offered is the one that the writer judges to represent the "best" usage.

academic aptitude. The combination of native and acquired abilities that are needed for school learning; likelihood of success in mastering academic work, as estimated from measures of the necessary abilities. (Also called *scholastic aptitude, school learning ability, academic potential*)

achievement test. A test that measures the extent to which a person has "achieved" something, acquired certain information, or mastered certain skills – usually as a result of planned instruction or training.

age norms. Originally, values representing typical or average performance for persons of various *age* groups; most current usage refers to sets of complete score interpretive data for appropriate successive age groups. Such norms are generally used in the interpretation of mental ability test scores.

alternate-form reliability. The closeness of correspondence, or correlation, between results on alternate (i.e., equivalent or parallel) forms of a test; thus, a measure of the extent to which the two forms are consistent or reliable in measuring whatever they do measure. The time interval between the two testings must be relatively short so that the examinees themselves are unchanged in the ability being measured. See RELIABILITY, RELIABILITY COEFFICIENT.

anecdotal record. A written description of an incident in an individual's behavior that is reported objectively and is considered significant for the understanding of the individual.

aptitude. A combination of abilities and other characteristics, whether native or acquired, that are indicative of an individual's ability to learn or to develop proficiency in some particular area if appropriate education or training is provided. Aptitude tests include those of general academic ability (commonly called mental ability or intelligence tests), those of special abilities, such as verbal, numerical, mechanical, or musical; tests assessing "readiness" for learning; and prognostic tests, which measure both ability and previous learning, and are used to predict future performance – usually in a specific field, such as foreign language, shorthand, or nursing.

Some would define "aptitude" in a more comprehensive sense. Thus, "musical aptitude" would refer to the combination not only of physical and mental characteristics but also of motivational factors, interest, and conceivably other characteristics, which are conducive to acquiring proficiency in the musical field.

arithmetic mean. A kind of average usually referred to as the *mean*. It is obtained by dividing the sum of a set of scores by their number.

average. A general term applied to the various measures of central tendency. The three most widely used averages are the arithmetic mean (mean), the median, and the mode. When the term "average" is used without designation as to type, the most likely assumption is that it is the *arithmetic mean*.

battery. A group of several tests standardized on the same sample population so that results on the several tests are comparable. (Sometimes loosely applied to any group of tests administered together, even though not standardized on the same subjects.) The most common test batteries are those of school achievement, which include subtests in the separate learning areas.

bivariate chart (bivariate distribution). A diagram in which a tally mark is made to show the scores of one individual on *two variables*. The intersection of lines determined by the horizontal and vertical scales form cells in which the tallies are placed. Such a plot provides frequencies for the two distributions, and portrays the relation between the two variables as a basis for computation of the product-moment correlation coefficient.

TM009 779

Issued by the Test Department, Harcourt Brace Jovanovich, Inc. HBJ

ceiling. The upper limit of ability that can be measured by a test. When an individual makes a score which is at or near the highest possible score, it is said that the test has too low a "ceiling" for him; he should be given a higher level of the test.

central tendency. A measure of central tendency provides a single most typical score as representative of a group of scores; the "trend" of a group of measures as indicated by some type of average, usually the *mean* or the *median*.

coefficient of correlation. A measure of the degree of relationship or "going-togetherness" between two sets of measures for the same group of individuals. The correlation coefficient most frequently used in test development and educational research is that known as the Pearson or *product-moment r*. Unless otherwise specified, "correlation" usually refers to this coefficient, but *rank*, *biserial*, *tetrachoric*, and other methods are used in special situations. Correlation coefficients range from .00, denoting a complete absence of relationship, to +1.00, and to −1.00, indicating perfect positive or perfect negative correspondence, respectively. See CORRELATION.

composite score. A score which combines several scores, usually by addition; often different weights are applied to the contributing scores to increase or decrease their importance in the composite. Most commonly, such scores are used for *predictive* purposes and the several weights are derived through multiple regression procedures.

concurrent validity. See VALIDITY (2).

construct validity. See VALIDITY (3).

content validity. See VALIDITY (1).

correction for guessing (correction for chance). A reduction in score for wrong answers, sometimes applied in scoring true-false or multiple-choice questions. Such scoring formulas ($R - W$ for tests with 2-option response, $R - \frac{1}{2}W$ for 3 options, $R - \frac{1}{3}W$ for 4, etc.) are intended to discourage guessing and to yield more accurate rankings of examinees in terms of their true knowledge. They are used much less today than in the early days of testing.

correlation. Relationship or "going-togetherness" between two sets of scores or measures; tendency of one score to vary concomitantly with the other, as the tendency of students of high IQ to be above average in reading ability. The existence of a strong relationship – i.e., a high correlation – between two variables does not necessarily indicate that one has any causal influence on the other. See COEFFICIENT OF CORRELATION.

criterion. A standard by which a test may be judged or evaluated; a set of scores, ratings, etc., that a test is designed to measure, to predict, or to correlate with. See VALIDITY.

criterion-referenced (content-referenced) test. Terms often used to describe tests designed to provide information on the specific knowledge or skills possessed by a student. Such tests usually cover relatively small units of content and are closely related to instruction. Their scores have meaning in terms of *what* the student knows or can do, rather than in their relation to the scores made by some external reference group.

criterion-related validity. See VALIDITY (2).

culture-fair test. So-called culture-fair tests attempt to provide an equal opportunity for success by persons of all cultures and life experiences. Their content must therefore be limited to that which is equally common to all cultures, or to material that is entirely unfamiliar and novel for all persons whatever their cultural background. See CULTURE-FREE TEST.

culture-free test. A test that is free of the impact of all cultural experiences; therefore, a measure reflecting only hereditary abilities. Since culture permeates all of man's environmental contacts, the construction of such a test would seem to be an impossibility. Cultural "bias" is not eliminated by the use of non-language or so-called performance tests, although it may be reduced in some instances. In terms of most of the purposes for which tests are used, the validity (value) of a "culture-free" test is questioned; a test designed to be equally applicable to all cultures may be of little or no practical value in any.

curricular validity. See VALIDITY (2).

decile. Any one of the nine points (scores) that divide a distribution into ten parts, each containing one-tenth of all the scores or cases; every tenth percentile. The first decile is the 10th percentile, the eighth decile the 80th percentile, etc.

deviation. The amount by which a score differs from some reference value, such as the mean, the norm, or the score on some other test.

deviation IQ (DIQ). An age-based index of general mental ability. It is based upon the difference or deviation between a person's score and the typical or average score for persons of his chronological age. Deviation IQs from most current scholastic aptitude measures are standard scores with a mean of 100 and a standard deviation of 16 for each defined age group.

diagnostic test. A test used to "diagnose" or analyze; that is, to locate an individual's specific areas of weakness or strength, to determine the nature of his weaknesses or deficiencies, and, wherever possible, to suggest their cause. Such a test yields measures of the components or subparts of some larger body of information or skill. Diagnostic achievement tests are most commonly prepared for the skill subjects.

difficulty value. An index which indicates the percent of some specified group, such as students of a given age or grade, who answer a test item correctly.

discriminating power. The ability of a test item to differentiate between persons possessing much or little of some trait.

discrimination index. An index which indicates the *discriminating power* of a test item. The most commonly used index is derived from the number passing the item in the highest 27 percent of the group (on total score) and the number passing in the lowest 27 percent.

distractor. Any incorrect choice (option) in a test item.

distribution (frequency distribution). A tabulation of the scores (or other attributes) of a group of individuals to show the number (frequency) of each score, or of those within the range of each interval.

equivalent form. Any of two or more forms of a test that are closely parallel with respect to the nature of the content and the number and difficulty of the items included, and that will yield very similar average scores and measures of variability for a given group. (Also referred to as *alternate, comparable,* or *parallel* form.)

error of measurement. See STANDARD ERROR OF MEASUREMENT.

expectancy table ("expected" achievement). A term with two common usages, related but with some difference:

(1) A table or other device for showing the relation between scores on a predictive test and some related outcome. The outcome, or criterion status, for individuals at each level of predictive score may be expressed as (a) an average on the outcome variable, (b) the percent of cases at successive levels, or (c) the probability of reaching given performance levels. Such tables are commonly used in making predictions of educational or job success.

(2) A table or chart providing for an interpretation of a student's obtained score on an achievement test with the score which would be "expected" for those at his grade level and with his level of scholastic aptitude. Such "expectancies" are based upon actual data from administration of the specified achievement and scholastic aptitude tests to the same student population. The term "anticipated" is also used to denote achievement as differentiated by level of "intellectual status."

extrapolation. In general, any process of estimating values of a variable beyond the range of available data. As applied to test norms, the process of extending a norm line into grade or age levels not tested in the standardization program, in order to permit interpretation of extreme scores. Since this extension is usually done graphically, considerable judgment is involved. Extrapolated values are thus to some extent arbitrary; for this and other reasons, they have limited meaning.

f. A symbol denoting the *frequency* of a given score or of the scores within an interval grouping.

face validity. See VALIDITY (1).

factor. In mental measurement, a hypothetical trait, ability, or component of ability that underlies and influences performance on two or more tests and hence causes scores on the tests to be correlated. The term "factor" strictly refers to a theoretical variable, derived by a process of *factor analysis* from a table of intercorrelations among tests. However, it is also used to denote the psychological interpretation given to the variable—i.e., the mental trait assumed to be represented by the variable, as verbal ability, numerical ability, etc.

factor analysis. Any of several methods of analyzing the intercorrelations among a set of variables such as test scores. Factor analysis attempts to account for the interrelationships in terms of some underlying "factors," preferably fewer in number than the original variables, and it reveals how much of the variation in each of the original measures arises from, or is associated with, each of the hypothetical factors. Factor analysis has contributed to an understanding of the organization or components of intelligence, aptitudes, and personality; and it has pointed the way to the development of "purer" tests of the several components.

forced-choice item. Broadly, any multiple-choice item in which the examinee is *required* to select one or more of the given choices. The term is most often used to denote a special type of multiple-choice item employed in personality tests in which the options are (1) of equal "preference value," i.e., chosen equally often by a typical group, and are (2) such that one of the options discriminates between persons high and low on the factor that this option measures, while the other options measure other factors. Thus, in the *Gordon Personal Profile,* each of four options represents one of the four personality traits measured by the *Profile,* and the examinee must select both the option which describes him *most* and the one which describes him *least.*

frequency distribution. See DISTRIBUTION.

g. Denotes *general* intellectual ability; one dimensional measure of "mind," as described by the British psychologist Spearman. A test of "*g*" serves as a general-purpose test of mental ability.

grade equivalent (GE). The grade level for which a given score is the real or estimated average. Grade-equivalent interpretation, most appropriate for elementary level achievement tests, expresses obtained scores in terms of *grade* and *month of grade,* assuming a 10-month school year (e.g., 5.7). Since such tests are usually standardized at only one (or two) point(s) within each grade, grade equivalents between points for which there are data-based scores must be "estimated" by *interpolation.* See EXTRAPOLATION, INTERPOLATION.

grade norms. Norms based upon the performance of pupils of given grade placement. See GRADE EQUIVALENT, NORMS, PERCENTILE RANK, STANINE.

group test. A test that may be administered to a number of individuals at the same time by one examiner.

individual test. A test that can be administered to only one person at a time, because of the nature of the test and/or the maturity level of the examinees.

intelligence quotient (IQ). Originally, an index of brightness expressed as the ratio of a person's mental age to his chronological age, MA/CA, multiplied by 100 to eliminate the decimal. (More precisely – and particularly for adult ages, at which mental growth is assumed to have ceased – the ratio of mental age to the mental age normal for chronological age.) This quotient IQ has been gradually replaced by the deviation IQ concept.

It is sometimes desired to give additional meaning to IQs by the use of verbal descriptions for the ranges in which they fall. Since the IQ scale is a continuous one, there can be no inflexible line of demarcation between such successive category labels as very superior, superior, above average, average, below average, etc.; any verbal classification system is therefore an arbitrary one. There appears to be, however, rather common use of the term *average* or *normal* to describe IQs from 90-109 inclusive.

An IQ is more definitely "interpreted" by noting the normal percent of IQs within a range which includes the IQ, and/or

[intelligence quotient (IQ), continued.]

by indicating its percentile rank or stanine in the total national norming sample. Column 2 of Table 1 shows the normal distribution of IQs for M = 100 and S.D. = 16, showing percentages within successive 10-point intervals. (For IQs whose S.D. is greater than 16, the percentages for the extreme IQ ranges will be larger, and those for IQs near the mean will be smaller, than those shown in the table.) Table 1 indicates that 47 percent, approximately one-half of "all" persons, have IQs in the 20-point range of 90 through 109; an IQ of 140 or above would be considered as extremely high, since fewer than one percent (0.6) of the total population reach this level, and fewer than one percent have IQs below 60. From the cumulative percents given in Column 3, it is noted that 3.1 percent have IQs below 70, usually considered the mentally retarded category. This column may be used to indicate the percentile rank (PR) of certain IQs. Thus an IQ of 119 has a PR of 89, since 89.4 percent of IQs are 119 or below; an IQ of 79 has a PR of 10.6, or 11. See DEVIATION IQ, MENTAL AGE.

Table 1. Normal Distribution of IQs with Mean of 100 and Standard Deviation of 16

(1) IQ Range	(2) Percent of Persons		(3) Cumulative Percent
140 and above	0.6		100.6
130-139	2.5		99.4
120-129	7.5		96.9
110-119	16.0		89.4
100-109	23.4	46.8	73.4
90- 99	23.4		50.0
80- 89	16.0		26.6
70- 79	7.5		10.6
60- 69	2.5		3.1
Below 60	0.6		0.6
Total	100.0		

internal consistency. Degree of relationship among the items of a test; consistency in content sampling. See SPLIT-HALF RELIABILITY.

interpolation. In general, any process of estimating intermediate values between two known points. As applied to test norms, it refers to the procedure used in assigning interpretive values (e.g., grade equivalents) to scores between the successive average scores actually obtained in the standardization process. Also, in reading norm tables it is necessary at times to interpolate to obtain a norm value for a score between two scores given in the table; e.g., in the table shown here, a percentile rank of 83 (from 81 + ⅓ of 6) would be assigned, by *interpolation*, to a score of 46; a score of 50 would correspond to a percentile rank of 94 (obtained as 87 + ⅗ of 10).

Score	Percentile Rank
51	97
48	87
45	81

inventory. A questionnaire or check list, usually in the form of a self-report, designed to elicit non-intellective information about an individual. Not tests in the usual sense, inventories are most often concerned with personality traits, interests, attitudes, problems, motivation, etc. See PERSONALITY TEST.

inventory test. An achievement test that attempts to cover rather thoroughly some relatively small unit of specific instruction or training. An inventory test, as the name suggests, is in the nature of a "stock-taking" of an individual's knowledge or skill, and is often administered prior to instruction.

item. A single question or exercise in a test.

item analysis. The process of evaluating single test items in respect to certain characteristics. It usually involves determining the difficulty value and the discriminating power of the item, and often its correlation with some external criterion.

Kuder-Richardson formula(s). Formulas for estimating the reliability of a test that are based on *inter-item consistency* and require only a single administration of the test. The one most used, formula 20, requires information based on the number of items in the test, the standard deviation of the total score, and the proportion of examinees passing each item. The Kuder-Richardson formulas are not appropriate for use with speeded tests.

mastery test. A test designed to determine whether a pupil has mastered a given unit of instruction or a single knowledge or skill; a test giving information on *what* a pupil knows, rather than on how his performance relates to that of some norm-reference group. Such tests are used in computer-assisted instruction, where their results are referred to as content- or criterion-referenced information.

mean (M). See ARITHMETIC MEAN.

median (Md). The middle score in a distribution or set of ranked scores; the point (score) that divides the group into two equal parts; the 50th percentile. Half of the scores are below the median and half above it, except when the median itself is one of the obtained scores.

mental age (MA). The age for which a given score on a mental ability test is average or normal. If the average score made by an unselected group of children 6 years, 10 months of age is 55, then a child making a score of 55 is said to have a mental age of 6–10. Since the mental age unit shrinks with increasing (chronological) age, MAs do not have a uniform interpretation throughout all ages. They are therefore most appropriately used at the early age levels where mental growth is relatively rapid.

modal-age norms. Achievement test norms that are based on the performance of pupils of normal age for their respective grades. Norms derived from such age restricted groups are free from the distorting influence of the scores of underage and overage pupils.

mode. The score or value that occurs most frequently in a distribution.

multiple-choice item. A test item in which the examinee's task is to choose the correct or best answer from several given answers or options.

N. The symbol commonly used to represent the number of cases in a group.

non-language test. See NON-VERBAL TEST.

non-verbal test. A test that does not require the use of words in the item or in the response to it. (Oral directions may be included in the formulation of the task.) A test cannot, however, be classified as non-verbal simply because it does not require reading on the part of the examinee. The use of non-verbal tasks cannot completely eliminate the effect of culture.

norm line. A smooth curve drawn to best fit (1) the plotted mean or median scores of successive age or grade groups, or (2) the successive percentile points for a single group.

normal distribution. A distribution of scores or measures that in graphic form has a distinctive bell-shaped appearance. Figures 1 and 2 show graphs of such a distribution, known as a *normal, normal probability,* or *Gaussian* curve. (Difference in shape is due to the different variability of the two distributions.) In such a normal distribution, scores or measures are distributed symmetrically about the mean, with as many cases up to various distances above the mean as down to equal distances below it. Cases are concentrated near the mean and decrease in frequency, according to a precise mathematical equation, the farther one departs from the mean. *Mean* and *median* are identical. The assumption that mental and psychological characteristics are distributed normally has been very useful in test development work.

norms. Statistics that supply a frame of reference by which meaning may be given to obtained test scores. Norms are based upon the actual performance of pupils of various grades or ages in the standardization group for the test. Since they represent average or typical performance, they should not be regarded as standards or as universally desirable levels of attainment. The most common types of norms are deviation IQ, percentile rank, grade equivalent, and stanine. Reference groups are usually those of specified age or grade.

objective test. A test made up of items for which correct responses may be set up in advance; scores are unaffected by the opinion or judgment of the scorer. Objective keys provide for scoring by clerks or by machine. Such a test is contrasted with a "subjective" test, such as the usual essay examination, to which different persons may assign different scores, ratings, or grades.

omnibus test. A test (1) in which items measuring a variety of mental operations are all combined into a single sequence rather than being grouped together by type of operation, and (2) from which only a single score is derived, rather than separate scores for each operation or function. Omnibus tests make for simplicity of administration, since one set of directions and one overall time limit usually suffice. The Elementary, Intermediate, and Advanced tests in the *Otis-Lennon Mental Ability Test* series are omnibus-type tests, as contrasted with the *Kuhlmann-Anderson Measure of Academic Potential,* in which the items measuring similar operations occur together, each with its own set of directions. In a *spiral-omnibus* test, the easiest items of each type are presented first, followed by the same succession of item types at a higher difficulty level, and so on in a rising spiral.

percentile (P). A point (score) in a distribution at or below which fall the percent of cases indicated by the percentile. Thus a score coinciding with the 35th percentile (P_{35}) is regarded as equaling or surpassing that of 35 percent of the persons in the group, and such that 65 percent of the performances exceed this score. "Percentile" has nothing to do with the percent of correct answers an examinee makes on a test.

percentile band. An interpretation of a test score which takes account of the measurement error that is involved. The range of such bands, most useful in portraying significant differences in battery profiles, is usually from one standard error of measurement below the obtained score to one standard error of measurement above it.

percentile rank (PR). The expression of an obtained test score in terms of its position within a group of 100 scores; the percentile rank of a score is the percent of scores equal to or lower than the given score in its own or in some external reference group.

performance test. A test involving some motor or manual response on the examinee's part, generally a manipulation of concrete equipment or materials. Usually *not* a paper-and-pencil test.

(1) A "performance" test of mental ability is one in which the role of language is excluded or minimized, and ability is assessed by what the examinee *does* rather than by what he says (or writes). Mazes, form boards, picture completion, and other types of items may be used. Examples include certain *Stanford-Binet* tasks, the Performance Scale of *Wechsler Intelligence Scale for Children, Arthur Point Scale of Performance Tests, Raven's Progressive Matrices.*

(2) "Performance" tests include measures of mechanical or manipulative ability where the task itself coincides with the objective of the measurement, as in the *Bennett Hand-Tool Dexterity Test.*

(3) The term "performance" is also used to denote a test that is actually a *work-sample;* in this sense it may include paper-and-pencil tests, as, for example, a test in bookkeeping, in shorthand, or in proofreading, where no materials other than paper and pencil may be required, and where the test response is identical with the behavior about which information is desired. *SRA Typing Skills* is such a test.

The use of the term "performance" to describe a type of test is not very precise and there are certain "gray areas." Perhaps one should think of "performance" tests as those on which the obtained differences among individuals may *not* be ascribed to differences in ability to use verbal symbols.

personality test. A test intended to measure one or more of the non-intellective aspects of an individual's mental or psychological make-up; an instrument designed to obtain information on the affective characteristics of an individual—emotional, motivational, attitudinal, etc. – as distinguished from his abilities. Personality tests include (1) the so-called *personality* and *adjustment inventories* (e.g., *Bernreuter Personality Inventory, Bell Adjustment Inventory, Edwards Personal Preference Schedule)* which seek to measure a person's status

[**personality test,** continued.]
on such traits as dominance, sociability, introversion, etc., by means of self-descriptive responses to a series of questions; (2) *rating scales* which call for rating, by one's self or another, the extent to which a subject possesses certain traits; and (3) *opinion or attitude inventories* (e.g., *Allport-Vernon-Lindzey Study of Values, Minnesota Teacher Attitude Inventory*). Some writers also classify interest, problem, and belief inventories as personality tests (e.g., *Kuder Preference Record, Mooney Problem Check List*). See PROJECTIVE TECHNIQUE.

power test. A test intended to measure level of performance unaffected by speed of response; hence one in which there is either no time limit or a very generous one. Items are usually arranged in order of increasing difficulty.

practice effect. The influence of previous experience with a test on a later administration of the same or a similar test; usually an increased familiarity with the directions, kinds of questions, etc. Practice effect is greatest when the interval between testings is short, when the content of the two tests is identical or very similar, and when the initial test-taking represents a relatively novel experience for the subjects.

predictive validity. See VALIDITY (2).

product-moment coefficient (r). Also known as the Pearson r. See COEFFICIENT OF CORRELATION.

profile. A graphic representation of the results on several tests, for either an individual or a group, when the results have been expressed in some uniform or comparable terms (standard scores, percentile ranks, grade equivalents, etc.). The profile method of presentation permits identification of areas of strength or weakness.

prognosis (prognostic) test. A test used to predict future success in a specific subject or field, as the *Pimsleur Language Aptitude Battery*.

projective technique (projective method). A method of personality study in which the subject responds as he chooses to a series of ambiguous stimuli such as ink blots, pictures, unfinished sentences, etc. It is assumed that under this free-response condition the subject "projects" manifestations of personality characteristics and organization that can, by suitable methods, be scored and interpreted to yield a description of his basic personality structure. The *Rorschach* (ink blot) *Technique*, the *Murray Thematic Apperception Test* and the *Machover Draw-a-Person Test* are commonly used projective methods.

quartile. One of three points that divide the cases in a distribution into four equal groups. The lower quartile (Q_1), or 25th percentile, sets off the lowest fourth of the group; the middle quartile (Q_2) is the same as the 50th percentile, or median, and divides the second fourth of cases from the third; and the third quartile (Q_3), or 75th percentile, sets off the top fourth.

r. See COEFFICIENT OF CORRELATION.

random sample. A sample of the members of some total population drawn in such a way that every member of the population has an equal chance of being included – that is, in a way that precludes the operation of bias or "selection." The purpose in using a sample free of bias is, of course, the requirement that the cases used be representative of the total population if findings for the sample are to be generalized to that population. In a *stratified* random sample, the drawing of cases is controlled in such a way that those chosen are "representative" also of specified subgroups of the total population. See REPRESENTATIVE SAMPLE.

range. For some specified group, the difference between the highest and the lowest obtained score on a test; thus a very rough measure of spread or variability, since it is based upon only two extreme scores. Range is also used in reference to the possible spread of measurement a test provides, which in most instances is the number of items in the test.

raw score. The first quantitative result obtained in scoring a test. Usually the number of right answers, number right minus some fraction of number wrong, time required for performance, number of errors, or similar direct, unconverted, uninterpreted measure.

readiness test. A test that measures the extent to which an individual has achieved a degree of maturity or acquired certain skills or information needed for successfully undertaking some new learning activity. Thus a *reading readiness* test indicates whether a child has reached a developmental stage where he may profitably begin formal reading instruction. *Readiness* tests are classified as *prognostic* tests.

recall item. A type of item that requires the examinee to supply the correct answer from his own memory or recollection, as contrasted with a *recognition item*, in which he need only identify the correct answer.

Columbus discovered America in the year _____

is a *recall* (or *completion*) item. See RECOGNITION ITEM.

recognition item. An item which requires the examinee to recognize or select the correct answer from among two or more given answers (options).

Columbus discovered America in

(a) *1425* (b) *1492* (c) *1520* (d) *1546*

is a *recognition* item.

regression effect. Tendency of a predicted score to be nearer to the mean of its distribution than the score from which it is predicted is to its mean. Because of the effects of regression, students making extremely high or extremely low scores on a test tend to make less extreme scores, i.e., closer to the mean, on a second administration of the same test or on some predicted measure.

reliability. The extent to which a test is consistent in measuring whatever it does measure; dependability, stability, trustworthiness, relative freedom from errors of measurement. Reliability is usually expressed by some form of *reliability coefficient* or by the *standard error of measurement* derived from it.

reliability coefficient. The coefficient of correlation between two forms of a test, between scores on two administrations of the same test, or between halves of a test, properly corrected. The three measure somewhat different aspects of reliability, but all are properly spoken of as reliability coefficients. See ALTERNATE-FORM RELIABILITY, SPLIT-HALF RELIABILITY COEFFICIENT, TEST-RETEST RELIABILITY COEFFICIENT, KUDER-RICHARDSON FORMULA(S).

representative sample. A sample that corresponds to or matches the population of which it is a sample with respect to characteristics important for the purposes under investigation. In an achievement test norm sample, such significant aspects might be the proportion of cases of each sex, from various types of schools, different geographical areas, the several socioeconomic levels, etc.

scholastic aptitude. See ACADEMIC APTITUDE.

skewed distribution. A distribution that departs from symmetry or balance around the mean, i.e., from normality. Scores pile up at one end and trail off at the other.

Spearman-Brown formula. A formula giving the relationship between the reliability of a test and its length. The formula permits estimation of the reliability of a test lengthened or shortened by any multiple, from the known reliability of a given test. Its most common application is the estimation of reliability of an entire test from the correlation between its two halves. See SPLIT-HALF RELIABILITY COEFFICIENT.

split-half reliability coefficient. A coefficient of reliability obtained by correlating scores on one half of a test with scores on the other half, and applying the Spearman-Brown formula to adjust for the doubled length of the total test. Generally, but not necessarily, the two halves consist of the odd-numbered and the even-numbered items. Split-half reliability coefficients are sometimes referred to as measures of the *internal consistency* of a test; they involve content sampling only, not stability over time. This type of reliability coefficient is inappropriate for tests in which speed is an important component.

standard deviation (S.D.). A measure of the variability or dispersion of a distribution of scores. The more the scores cluster around the mean, the smaller the standard deviation. For a normal distribution, approximately two thirds (68.3 percent) of the scores are within the range from one S.D. below the mean to one S.D. above the mean. Computation of the S.D. is based upon the square of the deviation of each score from the mean. The S.D. is sometimes called "sigma" and is represented by the symbol σ. (See Figure 1.)

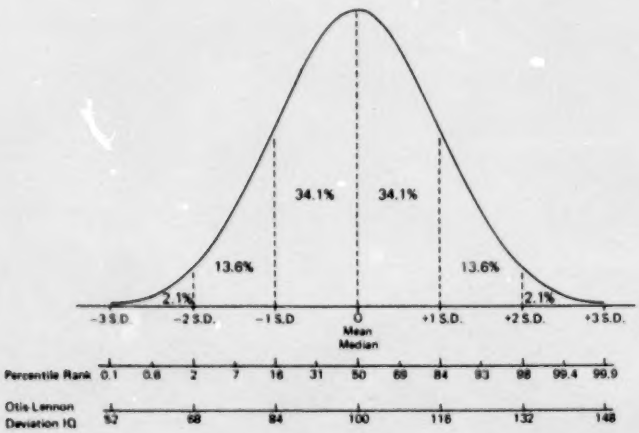

Figure 1. Normal curve, showing relations among standard deviation distance from mean, area (percentage of cases) between these points, percentile rank, and IQ from tests with an S.D. of 16.

standard error (S.E.). A statistic providing an estimate of the possible magnitude of "error" present in some obtained measure, whether (1) an *individual* score or (2) some *group* measure, as a mean or a correlation coefficient.

(1) standard error of measurement (S.E. Meas.): As applied to a single obtained score, the amount by which the score may differ from the hypothetical true score due to errors of measurement. The larger the S.E. Meas., the less reliable the score. The S.E. Meas. is an amount such that in about two-thirds of the cases the obtained score would not differ by more than one S.E. Meas. from the true score. (Theoretically, then, it can be said that the chances are 2:1 that the actual score is within a band extending from *true score minus 1 S.E. Meas.* to *true score plus 1 S.E. Meas.*; but since the true score can never be known, actual practice must reverse the true-obtained relation for an interpretation.) Other probabilities are noted under (2) below. See TRUE SCORE.

(2) standard error: When applied to group averages, standard deviations, correlation coefficients, etc., the S.E. provides an estimate of the "error" which may be involved. The group's size and the S.D. are the factors on which these standard errors are based. The same probability interpretation as for S.E. Meas. is made for the S.E.s of group measures, i.e., 2:1 (2 out of 3) for the 1 S.E. range, 19:1 (95 out of 100) for a 2 S.E. range, 99:1 (99 out of 100) for a 2.6 S.E. range.

standard score. A general term referring to any of a variety of "transformed" scores, in terms of which raw scores may be expressed for reasons of convenience, comparability, ease of interpretation, etc. The simplest type of standard score, known as a z-score, is an expression of the *deviation* of a score from the mean score of the group *in relation to* the standard deviation of the scores of the group. Thus:

$$\text{standard score (Z)} = \frac{\text{raw score (X)} - \text{mean (M)}}{\text{standard deviation (S.D.)}}$$

Adjustments may be made in this ratio so that a system of standard scores having any desired mean and standard deviation may be set up. The use of such standard scores does not affect the relative standing of the individuals in the group or change the shape of the original distribution. T-scores have a M of 50 and an S.D. of 10. Deviation IQs are standard scores with a M of 100 and some chosen S.D., most often 16; thus a raw score that is 1 S.D. above the M of its distribution would convert to a standard score (deviation IQ) of $100 + 16 = 116$. (See Figure 1.)

Standard scores are useful in expressing the raw scores of two forms of a test in comparable terms in instances where tryouts have shown that the two forms are not identical in difficulty; also, successive levels of a test may be linked to form a continuous standard-score scale, making across-battery comparisons possible.

standardized test (standard test). A test designed to provide a systematic sample of individual performance, administered according to prescribed directions, scored in conformance with definite rules, and interpreted in reference to certain normative information. Some would further restrict the usage of the term "standardized" to those tests for which the items have been chosen on the basis of experimental evaluation, and for which data on reliability and validity are provided. Others would add "commercially published" and/or "for general use."

stanine. One of the steps in a nine-point scale of standard scores. The stanine (short for *standard-nine*) scale has values from 1 to 9, with a mean of 5 and a standard deviation of 2. Each stanine (except 1 and 9) is ½ S.D. in width, with the middle (average) stanine of 5 extending from ¼ S.D. below to ¼ S.D. above the mean. (See Figure 2.)

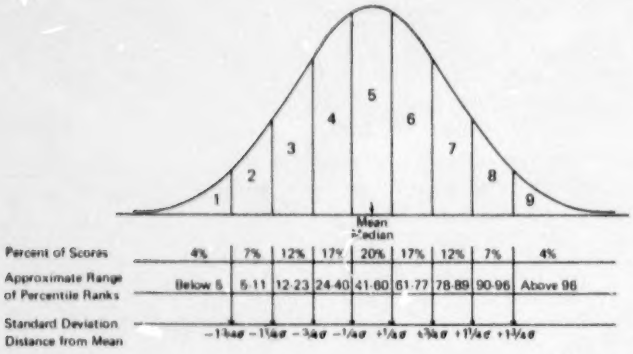

Figure 2. Stanines and the normal curve. Each stanine (except 1 and 9) is one half S.D. in width.

survey test. A test that measures general achievement in a given area, usually with the connotation that the test is intended to assess group status, rather than to yield precise measures of individual performance.

t. A critical ratio expressing the relationship of some measure (mean, correlation coefficient, difference, etc.) to its standard error. The size of this ratio is an indication of the significance of the measure. If *t* is as large as 1.96, significance at the .05 level is indicated; if as large as 2.58, at the .01 level. These levels indicate 95 or 99 chances out of 100, respectively.

taxonomy. An embodiment of the principles of classification; a survey, usually in outline form, such as a presentation of the objectives of education.

test-retest reliability coefficient. A type of reliability coefficient obtained by administering the same test a second time, after a short interval, and correlating the two sets of scores. "Same test" was originally understood to mean identical content, i.e., the same form; currently, however, the term "test-retest" is also used to describe the administration of different forms of the same test, in which case this reliability coefficient becomes the same as the alternate-form coefficient. In either case (1) fluctuations over time and in testing situation, and (2) any effect of the first test upon the second are involved. When the time interval between the two testings is considerable, as several months, a test-retest reliability coefficient reflects not only the consistency of measurement provided by the test, but also the stability of the examinee trait being measured.

true score. A score entirely free of error; hence, a hypothetical value that can never be obtained by testing, which always involves some measurement error. A "true" score may be thought of as the average score from an infinite number of measurements from the same or exactly equivalent tests, assuming no practice effect or change in the examinee during the testings. The standard deviation of this infinite number of "samplings" is known as the *standard error of measurement*.

validity. The extent to which a test does the job for which it is used. This definition is more satisfactory than the traditional "extent to which a test measures what it is supposed to measure," since the validity of a test is always specific to the purposes for which the test is used. The term validity, then, has different connotations for various types of tests and, thus, a different kind of validity evidence is appropriate for each.

(1) content, curricular validity. For achievement tests, validity is the extent to which the *content* of the test represents a balanced and adequate sampling of the outcomes (knowledge, skills, etc.) of the course or instructional program it is intended to cover. It is best evidenced by a comparison of the test content with courses of study, instructional materials, and statements of educational goals; and often by analysis of the processes required in making correct responses to the items. *Face validity*, referring to an observation of what a test appears to measure, is a non-technical type of evidence; apparent relevancy is, however, quite desirable.

(2) criterion-related validity. The extent to which scores on the test are in agreement with *(concurrent validity)* or predict *(predictive validity)* some given criterion measure. Predictive validity refers to the accuracy with which an aptitude, prognostic, or readiness test indicates future learning success in some area, as evidenced by correlations between scores on the test and future criterion measures of such success (e.g., the relation of score on an academic aptitude test administered in high school to grade point average over four years of college). In concurrent validity, no significant time interval elapses between administration of the test being validated and of the criterion measure. Such validity might be evidenced by *concurrent* measures of academic ability and of achievement, by the relation of a new test to one generally accepted as or known to be valid, or by the correlation between scores on a test and criteria measures which are valid but are less objective and more time-consuming to obtain than a test score would be.

(3) construct validity. The extent to which a test measures some relatively abstract psychological trait or construct; applicable in evaluating the validity of tests that have been constructed on the basis of an analysis (often factor analysis) of the nature of the trait and its manifestations. Tests of personality, verbal ability, mechanical aptitude, critical thinking, etc., are validated in terms of their construct and the relation of their scores to pertinent external data.

variability. The spread or dispersion of test scores, best indicated by their standard deviation.

variance. For a distribution, the average of the squared deviations from the mean; thus the square of the standard deviation.

TEST SERVICE NOTEBOOKS are issued from time to time as a professional service of the Test Department, Harcourt Brace Jovanovich, Inc. Inquiries, comments, or requests for additional copies may be addressed to the office nearest you. Write: Test Department, Harcourt Brace Jovanovich, Inc., New York 10017 • Chicago 60648 • San Francisco 94109 • Atlanta 30309 • Dallas 75235.